International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.141-148
http://dx.doi.org/10.14257/ijdta.2016.9.5.14

Cloud Data Migration Method Based on ABC Algorithm

[1]Geng Yushui, [2]Yuan Jiaheng and [3]Sun Tao
[1,2,3] *Qilu University of Technology*
gys@qlu.edu.cn, venusyjh@163.com, suntao0906@163.com

Abstract

Cloud storage systems has played an important role in the support of large-scale and high-performance cloud application. For the purposes of the cloud storage system, data migration is the key technology to the elastic load balancing.In this paper, we take data migration issues into the load-balancing scenarios and propose the data migration method based on the ABC algorithm. At the same time, we validated the method.In the method,we compute the load of each storage node through the comprehensive evaluation system.The system contains the following four aspects, including the available CPU,available memory, data access heat and the system response time.The cloud storage system performs the data migration operation according to the data obtained from the comprehensive evaluation system.The results show that this method can satisfy the application needs.At the same time, this method can reflect the integrated load of each node, and it also can achieve the optimal performance of the cloud storage system.

Keywords*: Cloud storage system, Data migration, ABC algorithm, Load balancing*

1. Introduction

Mass data storage with high reliability and scalability is a huge challenge for the internet companies while the traditional database is often hard to meet the demand.And the most retrieval for a specific system is based on the query of the primary key,in which case, the efficiency of relational database would be reduced, and the expansion would also become a big problem in the future.In this case, the use of cloud storage system would be a good choice. In terms of the deployment of cloud storage systems, data migration is a key technology to the node dynamic expansion and elastic load balancing.But a large number of state synchronization would bring many impact for system performance in the data migration process.Therefore, how to effectively reduce the migration cost is the biggest problems that cloud service provider should make efforts to solve. However, the state of the storage system, the new virtual environment, the user stringent latency requirements and unpredictability of access to the data have brought many new challenges for the data migration.

In this paper, we analyzed the background of data migration method, and put the data migration issues into the load-balancing scenarios.We use the load balancing framework of hadoop and propose a data migration method based on ABC algorithm. The innovation of this paper is that we apply ABC algorithm in data migration between different node of the distributed system. Experimental results show that the main role of ABC algorithm is to reduce the impact of data migration on system performance.At the same time,it also can increase system load balancing degree.

2. Related Work

The US National Institute of Standards and Technology (NIST) described the relevant concept of cloud computing and had given the architecture of cloud computing [8]. At present, cloud storage is a hot spot. Many scholars were studied about the concept of cloud storage system architecture and large data storage [2,6,12], however, the research

ISSN: 2005-4270 IJDTA

related to cloud storage data migrations is less. Julian Martin Kunkel discusseed load balancing mechanism of the parallel file system based on sub file migration [4]. Luoqiu Ming proposed a method of load balancing in the form of copies of data in "An Algorithm Independent PVFS Load Balancing Mechanism"[5]. The use of the copy can avoid changing the file allocation compartmentalization and also can solve the data moving back problems.Thereby, it can reduce system overhead. In the eighth literature, the author proposed a grid computing environment of mixed load balancing strategy, it greatly considered the network performance [11,13]. Chiu of Washington State University who studied data migration issues of the cache node, and put forward a data migration method based on greedy strategy [1]. Based on these research results, we propose a new method of data migration based on the ABC algorithm in the distributed system.Through the comprehensive assessment on node storage space, the CPU, memory space, and the heat of node access, a new kind of load balancing algorithm was provided, which can realize the migration and scheduling between the data storage node, so that it can improve the comprehensive performance of the entire cloud storage system.

3. Load Balancing Degree Evaluation Model

There are two types of imbalance in the cloud storage system: data skew and load skew. Since the cost of weight-balanced is high,the paper mainly focuses on the second type of imbalance, namely,load skew often causes node overload .At the same time, we assume that the system has a high tolerance to the other types of uneven.Assumed that the load capacity of nodes i is b_i . Firstly, we carry out non-dimensional treatment for each node based on the formula (1), B_i represent that the node i can withstand the maximum load.We have standardized the load for each node based on the formula (2).Assume p={p1,p2,p3,…pn} (n are the number of nodes)are the standardized node load. According to Shannon's theory, information entropy can measure the degree of order of the system.Here, we also use information entropy table to show the distribution of cluster load, and the calculation is shown in Equation (3) [9].

$$l_i = b_i / B_i \quad (1) \qquad p_i = l_i / \sum_{i=1}^{n} l_i \quad (2) \qquad H(p) = -\sum_{i=1}^{n} p_i * logp_i \quad (3)$$

Obviously, when pi=1/n (i = 1,2,3, ..., n), *i.e.*,the load of each node is equal, the system gets maximum entropy H (P) max = log (n). To better show load distribution,we constructed balance function T values ranging from 0 to 1,showed in equation (4). Balancing function is the ratio of the results calculated by equation (3) to the maximum entropy, namely normalized entropy.

$$T = -\sum_{i=1}^{n} p_i * logp_i / H(P)_{max} = \sum_{i=1}^{n} p_i * logp_i / log(1/n) \quad (4)$$

4. ABC Algorithm Model

4.1. Principle of ABC algorithm

In the basic ABC algorithms, artificial bee colony algorithm contains three kinds of individuals: employment bee, observed bee and scouts bee.Each employment bee corresponds to definite nectar. According to the abundance of nectar, the ABC algorithm use roulette way to hire the observed bee gather honey. If nectar source have no improvement after several update, they would give up the nectar, the employed bees turnes into scouts bee, then search for new nectar randomly.

4.2 Description of the Problem

When the load balancing of storage system is tilted, to return to equilibrium, the system needs to perform data migration.The value of the load balancing can be detected according to the certain entropy value.The higher of the entropy value indicate that the load distribution is more uniform. In this process, we can set a threshold.When the entropy reach this threshold, system performs the data migration operation. To improve the efficiency of data migration and reduce resource consumption of the system, we use artificial bee colony algorithm in the data migration process.

To more effectively identify hot data and minimize the amount of data migration, we use the concept of block HDFS used, the default size is 64M.The file of HDFS is divided into one or more memory blocks. Each block is an independent storage unit, the data is allocated in the form of blocks on a cluster server [10]. In this article, we put each block as a partition and the partition is the basic unit of data migration and load monitoring.

Before calling migration algorithm, all the storage nodes are grouped on In_set or Out_set according to the relationship between the overall system load and the number of node.In a cloud environment, the target of cloud data migration is to achieve a balance between the nodes through the load data migration, that is, the system must migrate some data of Move-out node to Move-in node. “For this type of problem”, in this paper, we use artificial bee colony algorithm to calculate the minimum consumption of each partition migration "to control global and local search" and ultimately the system achieves resource optimization.

4.3 Build Model

The ABC algorithm is applied to each node of Out_set, that is, each node is relatively independent in the migration process.

Table 1. Correspondence Table

ABC Algorithm	Data Migration Method
nectar source	Move-in node
beehive	Move-out node
Lead bee	Leading packet(The feedback message)
observed bee	data partition
scouts bee	Investigative packet
richness	fitness value

Explain:

(1) The leading packets carry the following information: location information of nectar, network bandwidth, the amount of data that the Move-in node can accept

(2) We assume that node i of In_set in three-dimensional space location as X_i=(x_{i1} x_{i2} , x_{i3}). To calculate the solutions we adopt the following fitness function (we assumed that equal importance to each element of X_i):

$$fit(X_i) = -x_{i1} + x_{i2} + x_{i3} \tag{5}$$

Explain: x_{i1} x_{i2} x_{i3} represent normalized data

$x_{i1} = \frac{d_{i1}}{\min\{d_{11}, d_{21}, \ldots, d_{N1}\}}$, d_{k1} represent the distance of data partition to node i.

$x_{i2} = \frac{b_{i1}}{\min\{b_{11}, b_{21}, \ldots, b_{N1}\}}$, b_{i1} represent real-time network bandwidth.

$x_{i3} = \frac{q_{i1}}{\min\{q_{11}, q_{21}, \cdots, q_{N1}\}}$, q_{i1} represent the amount of data that the Move-in node can accept.

(1) Node initialization

In the initial phase, the storage system is initialized according to the relationship between normalized load values of the system with 1/n.Finally, N node of In_set obtained by the initialization.N investigation messages correspond to N food source.In this case, then the investigative message become to the leading message, that is, a leading message corresponds to a node of In_set. In addition, each food source can be seen as an individual source in ABC algorithm, then, N individuals make up the population of the algorithm.

In the following stage, the node of Out_set selects a node in In_set to migrate data in a certain probability. Furthermore, for the node of In_set, the higher yields, the greater probability to be selected. If the leading messages found that the nodes no longer be updated after several successive iterations, they would give up the node.Then they should transform from leading messages to investigative message and continue to find new node of In_set.

(2)The formula of search new Move-in node

Recording the best value so far, and launched a search in the neighborhood of the current Move-in node, search for a new Move-in node accroding the formula:

$$v_{ij}^{t+1} = x_{ij}^{t} + rand[-1,1]_{ij}(x_{ij}^{t} - x_{kj}^{t}) \tag{6}$$

j∈{ 1, 2, … , D },k∈{ 1, 2, …, N }, $rand[-1,1]_{ij}$ is a random number between -1 to 1.The probability of Move-out node select the leading packet (in this case the leading packet contains information about the Move-in node):

$$P_i = \frac{fit(X_i)}{\sum_{n=1}^{N} fit(X_k)} \tag{7}$$

The richer of the Move-in node, the greater probability to be selected by the Move-out node.

(3) The investigation packets generated

To avoid plunging local optimum, when the Move-in node have no improvement after ‘limit’times update, the Move-in node would be gave up.And then the Move-in node is recorded in taboo table,then the leading packet corresponding to the Move-in node changes into the scout packet,in addition, randomly generates a new position instead of the original immigration node.

(4) Abandoned Move-in node

The leading message generates a new Move-in node V_i in the around of node X_i.If the richness of V_i is higher than X_i,then V_i replace X_i.

The ABC algorithm provides some regulations: if a food source is not improved within a predetermined number of iterations, the food source would be abandoned. Obviously, the preset number of iterations is a very important parameter, usually called "limit". The variables $trial_i$ can be used to record the number of times that the food source is updated, the calculation formula is defined as:

$$trial_i = \begin{cases} 0, & fit(X_i) < fit(V_i) \\ trial_i + 1, & fit(X_i) \geq fit(V_i) \end{cases} \tag{8}$$

That is, if the Move-in node X_i has not been improved after the limit updated, that is, max ($trial_i$)> limit, it shows that the resulting solution is the local optimal solution at this time. In this case, the leading packet corresponding to the Move-in node changes into Investigative packet, the Investigative packet use the escape operator calculate a new Move-in node Z, such as formula (9): $Z_i = X_{\min} + rand[-1,1] \times (X_{\max} - X_{\min})$ Among them,

$$Z_i = (z_{i1}, z_{i2}, z_{i3}) \quad (9)$$

4.4 The Process of ABC Algorithm

Parameter setting: the number of food sources and the number of leading packet is equal.If the leading packet found that the food source is abdonded, it changes into scouts packet. The current number of iterations is "$trial_i$".The number of parameters is the" *limit*".

Pseudo-code

Import：In_set ,Out_set Output：Current optimal solution

1.Initialsolution()
// Initialize the solution X_i,(i=1,2……N)
2.cycle=0;
3.repeat：
4.Hire bee phase：ProcessEmployedBee();
GenerateNeighborMemorySolution();
//Produce adjacent scheme
Determine whether to give up the move-in node;
Determine whether update the move-in node information;
5.Scouts phase：ProcessScoutBee();
//Search the move-in node around the move-out node ;
If it found the better move-in node ,the update the node information;
6.observed bee phase：Observe and choose a better node;
$trial_i = trial_i + 1$;
7.Record the best solution so far;
8.ntil the $trial_i$ is bigger than the maximum cycle times, out of the circulation
9.Returns the best solution, output the optimal result

Phase of the cycle is divided into the following six phase:

①Hire bee phase ② Calculate the probability of food source is selected by the bee

③ Follow bee phase ④ Scout bees phase

⑤ Recording the optimal food source location so far, that is , the optimal solution.

⑥$trial_i$ plus one,if $trial_i$ failure to reach the maximum number of iterations, then get to the next cycle stages

5. Experiment and Conclusions

5.1 Experimental Environment and Setting

The experiment environment includes three rack, as shown in Figure 1. The first rack contains one computer A. The second rack contains two kinds of computer B and C. The third rack contains one computer D. At the same time computer A as the client. Experimental topology shown in Figure 1.

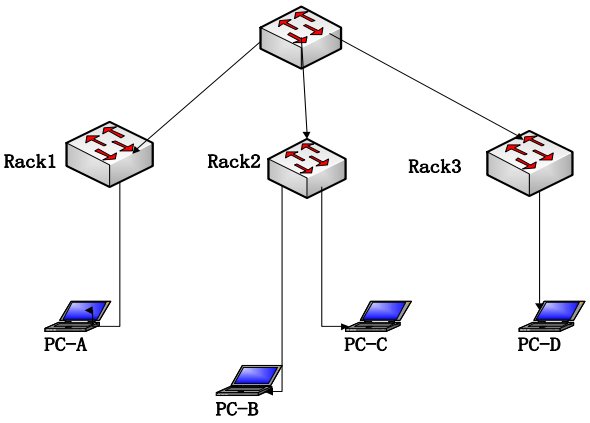


Figure 1. The Experimental Topology

5.2 The Results of Experiment

The node A store 1.7G data, the first copy is stored in the node A, the utilization rate of A is 100%, the second and third copies are stored in the B, C, D node, the utilization rate of B,C,D followed by 75%, 25%, 50%. Four nodes are not balanced. Here, the imbalance rate threshold is set to 0.1 [3,7,9]. Data migration method based ABC algorithm and the greedy algorithm through data on the balance after rounding decimal places in Table 2 after the storage state.

Table 2. Data Storage Rate Before and After the Balance Operation

Node	The configuration capacity(GB)	state befor the balance operation		After the data migration method based on ABC algorithm		After the data migration method based on greedy algorithm	
		usage amount(GB)	usage rate(%)	usage amount(GB)	usage rate(%)	usage amount(GB)	usage rate(%)
A	1	1.0	100	0.70	70	0.70	70
B	2.0	1.4	70	1.30	65	1.40	70
C	4.0	1.0	25	2.70	67.5	2.50	62.5
D	4.0	2.0	50	2.20	55	2.40	60
Standardization of entropy T		0.680098		0.892788		0.9203276	

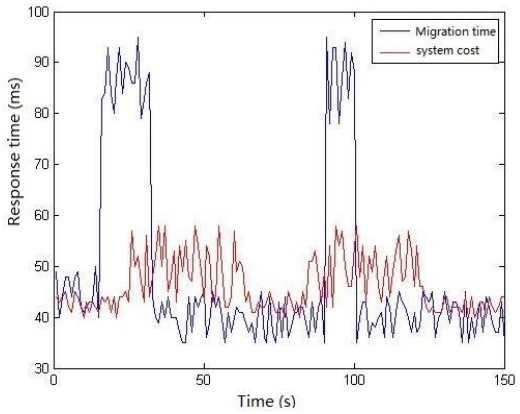


Figure 2. The Image Data of Data Migration Method Based on ABC Algorithm

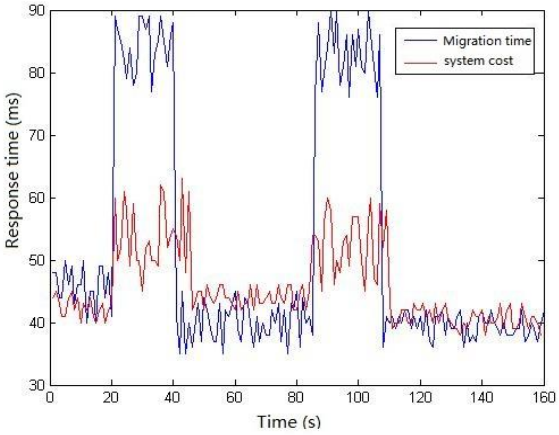


Figure 3. The Image Data of Data Migration Method Based on Greedy Algorithm

5.3 Conclusion

The experiment have respectively given the data figure of response time and system consuming about two kinds algorithms. The image shows that data migration method based on ABC algorithm can effectively reduce response time and system consumption. We can get the following conclusions from the experimental data, compared with the data migration method based on greedy algorithm, the data migration method based on ABC algorithm can quickly reduce the load tilt, and migration consumption is smaller. But when the entropy of the system gradually converge to the set threshold value, it may fall into local optimum.

5.4 Deficiencies

This method is not applicable to the system which produce sudden load continuously. Therefore, if the stability period is lower than the data migration time, the execution migration will introduce unnecessary expenses.The next work intends to consider the load stable period, we can use time series model and cost model of the system, then make improvement on the quality of the data migration strategy in the further.

In the experiments, in order to reduce the complexity of the issue, this article assume that the prediction model dose not consider CPU and memory heterogeneous situation.In the future, we will further study whether heterogeneous nodes occur that could impact the method.

Acknowledgements

Project supported by the projects of Shandong Province Higher Educational Science and Technology Program, China (No. J12LN20); Project supported by the projects of Shandong Province Science and Technology Development Plan, China (No. 2014GGX101052); Project supported by the projects of Shandong special independent innovation and achievements transformation, China(No. 2014ZZCX03408); Project supported by the projects of Shandong province natural science foundation, China (No. ZR2014FQ021).

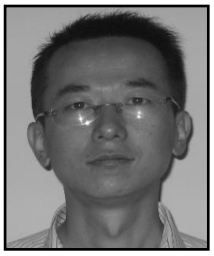
References

[1] D. Chiu, A. Shetty and G. Agrawal, "Elastic cloud caches for accelerating service-oriented computations", In: Proceeding of the ACM/IEEE Int'l Conference for High Performance Computing, Networking, Storage and Analysis, **(2010)**, pp. 1-11.
[2] D. Lang, "Cloud computing and cloud storage technology research", China new communication.
[3] G. Zhi, "Realization of the communication between devices of bacnet/ip and real-time database", Review of computer engineer studies, vol. 1, no. 1, **(2015)**, pp. 21-26.
[4] J. M. Kunkel, "Towards Automatic Load Balancing of a Parallel File System with Sub-file Based Migration", Philosophy, **(2007)**.
[5] L. Qiuming, "An independent PVFS load balancing mechanism algorithm", Computer engineering and Application, vol. 42, no. 19, **(2006)**, pp. 115-118.
[6] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica and M. Zaharia, "A view of cloud computing", Communications of the ACM, **(2010)**.
[7] M. Guo, S. L. Yang, H. Yan, L. F. Kan and B. Yang, "Mobile video alarm system based on cloud computing", Review of computer engineer studies, vol. 1, no. 2, **(2014)**, pp. 5-10.
[8] "NIST, Cloud computing reference architecture", NIST Special Publication, **(2011)**, pp. 500-292.
[9] Q. Xiulei, Z. Wenbo and W. Wei, "Cost sensitive data migration method for cloud Key / Value Storage System", Journal of Software, **(2013)**, pp. 1403-1417.
[10] W. Li, "Implementation and Simulation of geographic file system based on Hadoop", Peking University. **(2012)**.
[11] X. Luxin, "Research on the optimization of enrollment data resources based on cloud computing platform", Review of computer engineer studies, vol. 2, no. 2, **(2015)**, pp. 9-12.
[12] Y. Jiansha, "Storage architecture model based on cloud storage technology", Computer and Network, vol. 39, no. 7, **(2013)**, pp. 64-67.
[13] Y. Chuantao and C. Yong, "Mixed Load balancing strategy based on grid computing", Computer Engineering and Design, vol. 28, no. 16, **(2007)**, pp. 3925-3927.

Authors

Geng Yushui, He is a Professor of the Qilu University of Technology, Ph. D. supervisor, and commissioner of graduate students as well. He received his Master Degree in Shandong University (2006) and PhD in Tianjin University (2013).He has been interested in computer network, manufacturing information, cloud computing and big data.

Yuan Jiaheng, He is a graduate student of the Qilu University of Technology. He received the bachelor's degree in mathematics (2014). During the graduate, he has been researching computer network, manufacturing information, cloud computing and big data.



Sun Tao, He is a PhD, associate professor, his current research interests focus on big data, data integration and cloud computing.

